Non-Emergency Notification Timing for Drivers Doing Non-Driving-Related Tasks in Autonomous Vehicles: An Interruptibility Study

Hongyu Howie Wang
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
howiewang@cmu.edu

Jiya Gupta
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jiyag@andrew.cmu.edu

Nikolas Martelaro
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nikmart@cmu.edu

Abstract

Future high-level autonomous vehicles (AVs) will enable drivers to engage in non-driving-related tasks (NDRTs) during autopilot. Occasionally, an in-vehicle agent may need to notify drivers of important, yet not urgent, information. Through a four-session interruptibility study on a desktop autonomous driving simulator, we investigated how drivers assess their availability to receive notifications by rating moments as good or bad for interruption. Our results suggest drivers fall into four notification availability groups: always available, prioritizing NDRTs, task-content dependent, and mental-state dependent. Using multimodal behavioral data of the participants and vehicle data from the simulation, we trained a proof-of-concept classification model to determine the appropriate timing to send non-emergency notifications to drivers doing NDRTs. Head pose and gaze direction data from the eye tracker were crucial in the predictions. Based on our quantitative modeling and qualitative observation, we discuss the feasibility of notification timing prediction in the real world and design considerations from individual, task, and context perspectives.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; **Human computer interaction (HCI)**.

Keywords

Interruptions, Notification, Automotive, Eye tracking, Timing, Interruptibility, Autonomous vehicles, Driving simulator, Predictive models

ACM Reference Format:
Hongyu Howie Wang, Jiya Gupta, and Nikolas Martelaro. 2025. Non-Emergency Notification Timing for Drivers Doing Non-Driving-Related Tasks in Autonomous Vehicles: An Interruptibility Study. In *17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '25), September 21–25, 2025, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3744333.3747831

AutomotiveUI '25, Brisbane, QLD, Australia

ACM ISBN 979-8-4007-2013-0/25/09
https://doi.org/10.1145/3744333.3747831

1 Introduction

High-level autonomous vehicles (AVs) that require little to no supervision or intervention from human drivers (i.e., Level 3 and above on the SAE scale [59]) are becoming a reality. In the coming decades, many people may be fortunate enough to have access to autonomous vehicles for their daily commute. While the vehicle is in autopilot mode and cruising steadily, the driver may pull out their phone to read news and watch videos, or they may eat, sleep, window-gaze, apply cosmetics, or other non-driving related tasks (NDRTs) [8, 18, 45, 53].

Although autonomous driving allows drivers to engage in NDRTs with minimal concern, occasional notifications from in-car agents may still be necessary. UX researchers, motivated by Matthews et al. [43], have categorized such notifications into priority levels [40, 57], including *Demand-action* and *interrupt*, which require immediate attention; *make-aware*, which invites delayed actions; *change-blind*, offering optional, long-term information; and *ignore*, unimportant information requiring little attention. Non-emergency notifications like *make-aware* and *change-blind* must avoid being as intrusive as urgent ones, as overly frequent or exaggerated alerts can desensitize, disturb, or startle the driver [41, 50, 65], leading to dissatisfaction [48]. Burnett et al. [8] found that even after multiple days of exposure, participants could still be startled by takeover requests issued 60 seconds before the takeover. Delivering non-emergency notifications at appropriate moments enhances driver satisfaction and ensures effective communication.

In light of this, we investigate the timing of in-vehicle notifications in AVs. We propose that by understanding the driver's behavioral status, we can better predict when to push non-emergency notifications smoothly, timely, and noticeably. Our research questions are:

RQ1 When are drivers available for non-emergency notifications while doing non-driving-related tasks (NDRTs)?

RQ2 Is it possible for an in-vehicle agent to model and predict opportune moments to push non-emergency notifications to a driver doing NDRTs?

RQ3 What are some considerations for timing non-emergency notifications in a high-level autonomous car?

Our work builds on the "Is Now a Good Time?" project by Semmens et al. [61], which explored optimal timing for vehicle-driver communication of non-emergency information. We adapt their methods to study user experience in autonomous vehicles (AVs)

where occupants focus on non-driving tasks. We conducted a study using a desktop autonomous driving simulator to collect data on user behavior, vehicle status, and preferences for notification cues. Over four days, 22 participants experienced a simulated self-driving commute while engaging in tablet-based non-driving activities. We recorded their behavior using eye-tracking glasses, tablet sensors, and video cameras. Participants labeled moments as "good" or "bad" for receiving non-emergency notifications when prompted by an audio cue. The dataset includes approximately 700 labels alongside real-time eye tracking, tablet activity, video, and vehicle status data.

In the post-session survey, participants explained when they were available for notifications. Based on the responses, we categorized them into four categories, addressing RQ.1. This further motivated the development of a proof-of-concept model that predicts driver availability for non-emergency notifications. The model utilized eye gaze, head pose, tablet activity, and vehicle status within a time window, significantly outperforming chance and answering RQ.2. Insights from classifier training, qualitative observations, and data analysis informed us about notification timing design for high-level AVs, answering RQ.3.

Our key contributions are: 1) a procedure for notification timing design in a simulated AV environment; 2) a proof-of-concept classifier to predict appropriate moments for non-emergency notifications in AVs with passengers doing various NDRTs; 3) design considerations for timing the delivery of non-urgent content in future AVs where occupants engage in NDRTs.

2 Background

In this project, we simulated the AV experience on a stationary desktop driving simulator where participants engaged in NDRTs. This section highlights existing techniques that induce user engagement and estimate their cognitive load in an automotive environment. We also briefly discuss the concept of user interruptibility prediction.

2.1 Surrogate Tasks

Previous studies on driving distractions have incorporated a variety of secondary tasks such as visual search (ISO14198 [25]) to mimic demanding manual-visual activities [14, 56, 69] and *n*-back tests to challenge drivers' working memory capacity [9, 37, 44]. Other surrogate tasks to induce cognitive load include working memory games on a tablet [22] and complex sentence comprehension [15]. Some researchers preferred more natural NDRTs over repetitive tasks that intensively target cognitive load. For instance, Miller et al. [45] used video watching and reading as secondary tasks besides supervising the vehicle. Zeeb et al. [72] designed a task routine that included checking emails, reading news, and watching videos. Additional everyday tasks used as NDRTs (such as gaming and talking on the phone) are discussed in the review by Naujoks et al. [47]. For our study, we choose more organic activities closer to real-life NDRTs in order to elicit higher engagement and obtain more natural responses.

2.2 Cognitive load measurement

We aim to use cognitive load to confirm our manipulation of user engagement in NDRTs, estimate their **availability** for notifications, and gauge their **ability** to notice notifications. Cognitive load measurement can be conducted physiologically in real time through heart rate variation [39], pupil dilation [11], and skin temperature differences [1]. An intermittent and less invasive measurement technique is the detection response task [24] (also standardized in ISO17488 [33] for automotive use), where individuals respond to a visual, audio, or tactile signal. Reaction time has been shown to correlate with the cognitive demand for parallel tasks [63]. In addition to objective measurements, subjective cognitive load can be reported by individuals using questionnaires such as the NASA TLX [27], as used by Müller et al. [46] and Ko et al. [37] to confirm the driver's workload during NDRT. Cognitive load also often intertwines with situational awareness (SA) [19, 66]. Since a high cognitive load can impair SA in automotive contexts [4], assessing situational awareness can offer insight into the mental resources an activity demands. In this study, we include the detection response time, NASA TLX ratings, SA measurements, and pupil dilation as cognitive load indicators.

2.3 Studying Interruptibility

Studying notification timing involves understanding interruptibility, a topic explored in HCI for over two decades. In 2003, Hudson et al. [31] coded video and audio recordings of office workers. This enabled them to simulate an ideal scenario of ubiquitous computing, where sensors could provide comprehensive behavioral measurements. The workers were randomly prompted to rate their interruptibility during data collection. Using user-labeled data, the authors trained a classifier that achieved more than 75% accuracy.

Modern computing and sensing technologies have allowed researchers to collect various sensor data without manual coding. For mobile devices, Fischer [23] conducted a series of experiments exploring opportune interruptions using content, usage, and location data. Pielot et al. [54] trained a classification model that predicted user engagement with notifications based on phone usage data moments before. In human-robot interaction, Banerjee et al. [3] proposed a system for robots to gauge human interruptibility using social and contextual cues. In a recent automotive study named "Is Now a Good Time?", Semmens et al. [61] randomly asked drivers in a car if it was a good moment for them to receive non-driving information. The "yes/no" responses were used to label the data stream, including video recordings, physiological measurements, and vehicle CAN bus outputs. Wu et al. [68] used the dataset to train a deep neural network to predict "good" times for notifications. Our work draws inspiration from the "Is Now a Good Time?" project and extends the study of interruptibility to the context of AVs, focusing on notifications when drivers are not driving and doing NDRTs.

3 Data Collection Method

In this study, we collect behavioral data labeled by drivers doing NDRTs on our driving simulator. We combine different cognitive load induction and measurement techniques while preserving some organic interaction during the autonomous driving simulation. This section describes the simulator setup and data collection approach.

3.1 Driving Simulator and Data Collection Setup

Our driving simulator was based on a modified version of the video game *Grand Theft Auto V* running on a desktop computer. Through

Figure 1: Driving simulator setup. User activities during the simulated commute in an autonomous vehicle: (from left to right) (a) manual driving, (b) playing a mobile game on the tablet, (c) watching video lectures on the tablet, and (d) reading short stories on the tablet.

a combination of scripts and modifications developed by ourselves and the online community[1], we utilized the pre-programmed driving behavior of non-player characters to simulate autonomous driving. The vehicle "autonomously" drives to a set destination, following traffic rules and providing steering wheel feedback. The driving route formed a clockwise loop in the west and north parts of the map and covers city, highway, town, rural, and suburban driving, lasting approximately 25 minutes. Details of the route design can be found in Appendix A.1.

Multiple video cameras captured the front, side, and rear views of the driver: one above the computer monitor, two under the left and right corners of the desk, and one behind the driver on the left. A Tobii Pro Glasses 2 eye tracker tracked the driver's eye gaze and head pose and recorded the first-person-view video. A Pixel Android tablet registered user touches and device accelerometer readings, and the simulation program logged the status of the virtual environment (vehicle stats, cars and pedestrians nearby, etc.). A pair of LED light bars provided consistent lighting (for computer vision processing in future research), and a set of 5.1 speakers offered immersive audio similar to real-world driving. The complete setup is shown in Figure 1.

3.2 Data Labeling Procedure

Approved by our Institutional Review Board (#2023_00000224), our data labeling procedure was in the format of a longitudinal study, as we aim to encompass a range of NDRT engagement and reduce the novelty effect. On four different days, each participant experienced the same simulated commute route (with small differences in commute time and surrounding vehicles generated in-game). For each commute session, participants began with brief manual driving. They transitioned to autopilot mode by pressing a button on the steering wheel upon reaching the highway. During autonomous driving, participants performed an assigned non-driving-related task (NDRT), detailed in the next section. A looping audio signal ("boop" sound[2] every 3 seconds) was randomly triggered every 1 to 5 minutes. The signal volume increased until the participant *verbally* labeled the moment as "good/yes" or "bad/no" for receiving non-emergency notifications. Researchers logged these labels as data points. This procedure was repeated multiple times (typically >6) per session until the simulated vehicle reached its destination. Figure A.2 visually illustrates the process.

Following Semmens et al. [61], we chose to collect binary labels instead of numerical ratings of appropriateness. From the participants' perspective, this enables faster responses and reduces subjective variations in interpretations of a rating scale [58]. Practically, participants may have to think longer to determine how available the are on a scale and different participants may have different interpretation of more or less availability, reducing the reliability of their answers and making modeling more challenging. From a real-world system's perspective, the decision to push non-emergency notifications at a given moment is binary: to minimize annoyance, they should only be sent when the user is clearly available.

The audio signal served as a stimulus for the participant to label the moments, rather than representing an actual notification sound (an important but separate research topic). This procedure resembles an auditory detection response task used by Stojmenova et al.[63], but with two differences: 1) our audio stimulus gradually increased in volume to gauge participants' sensory threshold, which reflects both the ability to detect the signal and the cognitive load; it also ensures that we eventually get a response from the participant. 2) Participants spoke the label instead of pressing buttons, as the latter would have interfered with highly engaging NDRTs.

Consent was obtained on the first day of the study after the participant went through the researcher's verbal introduction and a practice run on a test track. Before starting each drive, the researcher restated the purpose of the study and the data labeling approach according to a script (Appendix A.3).

Participants were offered $10 for each session attended, and $10 more as a bonus for completing all four sessions. For each of the second, third, and fourth sessions, participants would also receive up to $10 depending on their performance in the NDRTs (see Section 3.3). In total, a participant would be compensated $50 to $80.

3.3 Non-Driving-Related Task (NDRT) Design

Table 1 provides details of NDRTs in the four sessions. In the introductory session, participants were introduced to the simulator and encouraged to "bring whatever they think they would interact with in a future autonomous vehicle." Following [8], this session observed participant choices and familiarized them with simulated

[1]At the time of writing, certain modification packages have been unlisted or archived by the original authors. The list of the mods used by us can be provided on request.

[2]The "Windows Balloon.wav" sound from Windows 10 operating system.

Table 1: Non-driving-related tasks (NDRTs) for different sessions. The Intro scenario is for the first session only. The order of Game, Video, and Reading scenarios is randomized.

Scenario	1 – Intro	2 – Game	3 – Video	4 – Reading
NDRT Type	Driver's own choice	Active, high cognitive load, high engagement	Passive, med. cognitive load, med. engagement	Passive, High cognitive load, high engagement
Surrogate NDRT	_	Playing a tablet game for monetary reward	Watching video lectures and completing a quiz for monetary reward	Reading short stories and completing quizzes for monetary reward

steady-state autonomous driving, reducing novelty effect. In subsequent sessions, participants engaged in one of three tasks on the Android tablet: playing the mobile game *Temple Run*, watching video lectures on culinary history, or reading sci-fi short stories. These tasks, inspired by previous NDRT research in AVs [45, 60, 72], varied in cognitive and physical demand. To incentivize engagement, we offered participants bonus compensation: $1 for every 5000m run in the game and $1 for each correct answer in 10-point quizzes about the lectures or short stories.

3.4 Situational Awareness Task and Post Session Survey

To confirm participants' cognitive load and NDRT engagement, we incorporated a secondary object identification task to measure situational awareness, as it can decrease with increased cognitive load during NDRTs [20]. In each session, three in-game objects were randomly placed in the course (see Appendix A.4). Participants would either verbally confirm seeing these objects during the drive or identify them in the post-session survey. This task was inspired by the Daze platform for testing SA in self-driving cars [62].

After the vehicle reached its destination, participants completed a survey (and quizzes for Reading and Video scenarios). The survey included optional demographic questions and 1–7 scale ratings on immersive tendencies, perceived presence [67], and task load (using NASA TLX [27]). The survey also confirmed objects identified by the participants. In addition, participants were asked to explain their interpretations of when was a "good" or "bad" moment for them during the session and to give suggestions about "less obtrusive" in-vehicle notifications.

3.5 Collected Metrics

For each session, we obtained time-stamped data, including vehicle status, video and audio recordings, eye gaze and head pose, tablet activities, and labels provided by the participants. We extracted data streams from 20 seconds before each signal went off to the moment when the signal was played (i.e., the onset of the first "boop;" $[t - 20, t]$) and from the moment to 20 seconds after ($[t, t + 20]$).

The survey responses included participants' interpretations of when a "good" or "bad" moment was, ratings of the simulation experience, ratings of the NDRT effort, SA object identification, and user feedback.

Drivers' ability to perceive cues is influenced by sensory and cognitive workload, which, in turn, impacts notification timing. To study this, we reviewed video and audio recordings and manually timed participants' signal detection and response. Detection was defined as the moment a participant looked up or made an initial utterance (e.g., smacking lips or saying "umm"), while response referred to when a participant provided a "good" or "bad" answer. From this, we derived four time measurements: 1) *Detection time*: Time from the first signal loop ("boop") to detection, directly related to the signal volume upon detection. 2) *Reaction time*: Time from the most recent "boop" to detection. 3) *Response time*: Time from the first "boop" to the participant's response. 4) *Decision time*: Time between detection and response. This may often be 0 due to familiarity or task-related urgency, though hesitation occasionally occurs. Figure A.4 illustrates these four measurements.

4 Data Collection Results

In this section, we report the immediate quantitative and qualitative results of our data collection.

4.1 Overview of Label Distribution

We recruited 22 participants (aged 18 to 40 with 0 to 22 years of driving experience, all having valid driver's licenses) from our university community. We obtained 86 valid sessions of data [3] (2 sessions were discarded due to equipment failure). This amounts to 700 instances of audio signal detection and labeling. Overall, we collected 444 (63.4%) "good" labels and 256 "bad" labels. Individuals on average rated 64.0% "good" (SD=32.2%) but ranged between 0% "good" and 100% "good" depending on personal preferences or scenario type. Three participants always answered "good" whenever prompted. Removing their data reduced the sample size to 603 instances. In this case, 347 (57.5% of 603) "good" labels remained. The entire 700 samples is our primary reference for study validation and qualitative observations, but the trimmed, 603-sample dataset is used for our modeling in Section 5 (with reasoning explained). The bar plots in Figures 2 display the ratio of "good" to "bad" labels in each session. Figure B.1 in Appendix B.1 shows the aggregated ratio of the labels in each scenario.

4.2 Simulation Immersiveness and Task Differences

Overall, the simulation environment offered adequate immersiveness according to participants' responses to the perceived presence questionnaire (Cronbach's $\alpha \approx 0.74$). The assigned NDRTs manipulated user engagement and cognitive load. Readers may refer to Appendix B.2 for a more in-depth statistical report.

The NASA TLX ratings show that Game ($M = 28.72$, $SD = 5.22$) and Reading ($M = 24.86$, $SD = 5.46$) induced significantly higher overall task load than Video ($M = 19.91$, $SD = 4.40$) (repeated measures ANOVA: $F(3, 63) = 21.81$, $p < 0.001$; posthoc Tukey's HSD: $p < 0.001$ for Game vs. Video and $p = 0.014$ for Reading vs.

[3]The processed data (excluding videos revealing the participants' identities) are publicly available as part of the artifact collection at: https://doi.org/10.1184/R1/c.7894613.

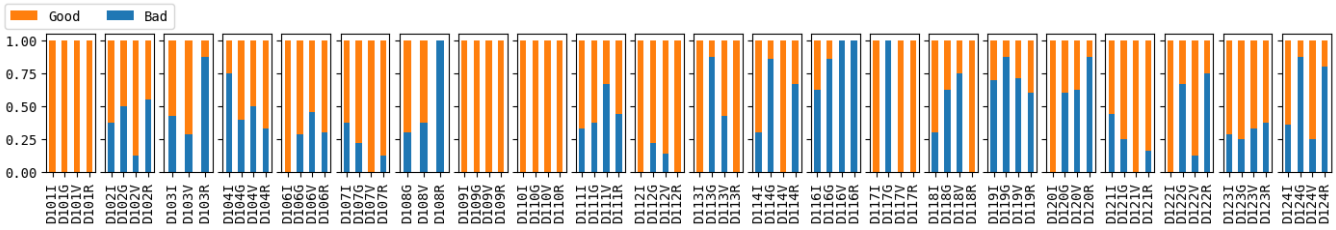


Figure 2: Relative frequency of labels in individual sessions. Each participant was assigned an alphanumeric identifier (D1xx). "I" denotes the Intro scenario, "G" the Game scenario, "V" the Video scenario, and "R" the Reading scenario. Some participants always gave out "good" labels in all scenarios while a few always said "bad" in one or two scenarios. 2 sessions involving D103 and D108 were removed due to data loss.

Video). The differences between Game and Reading in task load were significant only in physical demands (Tukey's HSD: $p = 0.010$ for Game $M = 3.36$, $SD = 1.99$ vs. Reading $M = 1.95$, $SD = 1.43$) and time demands (Tukey's HSD: $p = 0.026$ for Game $M = 5.82$, $SD = 1.37$ vs. Reading $M = 4.41$, $SD = 1.76$).

Two-way ANOVA revealed significant effects of both the number of simulator days and the NDRT scenario on detection time (day: $F(3, 691) = 4.02$, $p = 0.008$, scenario: $F(3, 691) = 18.96$, $p < 0.001$), response time (day: $F(3, 691) = 4.39$, $p = 0.005$, scenario: $F(3, 691) = 29.89$, $p < 0.001$), and decision time (day: $F(3, 691) = 6.05$, $p < 0.001$, scenario: $F(3, 691) = 18.96$, $p < 0.001$). Detection time was longer in the Video scenario ($M = 5.97$s, $SD = 3.22$s, Tukey's HSD $p < 0.001$) compared to both Game ($M = 4.10$s, $SD = 2.37$s) and Reading ($M = 3.88$s, $SD = 2.49$s).

Despite uncontrolled factors like graphics and lighting, situational awareness and pupil dilation measurements also reflected the manipulated task load and engagement. Participants in the Game scenario only identified $M = 0.43$ random objects on average ($SD = 0.79$), significantly less than in Video ($M = 1.35$, $SD = 1.07$, Tukey's HSD: $p = 0.009$). Participants playing the mobile game exhibited significantly larger pupil dilation ($M = 4.94$mm, $SD = 0.72$mm) (Tukey HSD $p < 0.001$ compared to both Video with $M = 3.73$mm, $SD = 0.55$mm and Reading with $M = 3.55$, $SD = 0.53$).

4.3 Qualitative Observations

We observed that participants opted for various activities in the introductory session. Meanwhile, participants provided different rationales for their "good/bad" responses and exhibited physical behavior patterns.

4.3.1 Activities in the Intro Scenario. During the Intro session, 21 out of 22 participants brought mobile devices (phone, tablet, or laptop); the remaining participant gazed at the windshield throughout the drive. 12 participants texted or messaged, 8 engaged in intellectual activities like studying, coding, or solving puzzles, and 6 performed light browsing or web searching. Furthermore, 6 participants read or studied text-based materials. Watching videos, task planning, and gaming each occurred twice. One participant took a piano lesson on their phone for some time during the drive.

4.3.2 Categories of Participants. To directly answer RQ.1, after each session, we asked participants in our survey, "In this session, how did you determine if it was a "good" or a "bad" moment to receive non-emergency notifications?" We then grouped responses based on similarity and found four categories of rationale that people provided for their choices. Please refer to Table B.1 for examples statements representing these categories:

(1) *Always available*: Participants found no conflict between the audio signal and their NDRT.
(2) *Prioritizing non-driving-related tasks*: If actively engaged in an NDRT, it was a bad moment; otherwise, good. Notification timing could be adjusted by monitoring driver actions and posture using a binary approach.
(3) *Task content dependent*: Challenging phases of NDRTs were bad moments; otherwise, good. Timing could be refined by evaluating task difficulty using a threshold-based approach.
(4) *Mental state dependent*: If the participant was "tired" or "in the zone," it would be a bad moment. Although more complex and implicit, timing may be estimated based on cognitive load, stress, or drowsiness.

4.3.3 Physical Behavior. We saw commonalities among the participants through the video recordings. In the Game, Video, and Reading scenarios, many participants hunched down and lowered the tablet when they had to focus on the content for a prolonged period (see Figure B.2 in Appendix B.4). When the participants were between rounds of the game (Figure B.3) or taking a break from the reading materials, they would slightly change their postures to a more upright position or look up briefly. Nuances also existed. For example, a few participants the Video scenario occasionally glanced up at the road before giving a "bad" label, likely due to their audio channel being heavily engaged with the video lecture. However, in general, we observed that "good" responses and signal detection times appeared to correlate with head pose, gaze direction, and tablet activity. Our observation further motivates us to explore the possibility of using machine-learning models to predict interruptibility based on these behavioral indicators.

4.3.4 User Suggestions. In the survey, participants provided suggestions for less obtrusive in-vehicle notification cues. D111 and D114 proposed triggering notifications when looking up to check road conditions. D112 recommended ambient notification lighting, while D101 and D119 suggested vibrotactile actuators. Some participants, like D103 and D113, expressed interest in previewing notification content to tailor their responses based on its importance.

5 Modeling

We attempt to answer RQ.2 by training a proof-of-concept classifier using the collected behavioral data and labels. Similar to Semmens et al. [61] and Wu et al. [68], we hypothesize that for drivers doing NDRTs, a machine learning classifier can identify good and bad moments for notifications with real-time behavioral measurements.

As noted in Section 4.1, the dataset used for the majority of our modeling excludes the three participants who always provided "good" labels. We argue that these participants would not directly benefit from a predictive model because they were always available for notifications *regardless* of activity or cognitive status. In practice, the "all good" users could opt out of notification management services; in model training, as the model recognized these users, it would yield many more correct "good" moment predictions, inflating the model's precision. Section 5.1.9 briefly reports the model performance in this case, which may be deemed less generalizable.

We aimed to focus on participants who *did not always* answer "good." Data from participants who reported exclusively "good" or "bad" only during specific NDRTs remained included: the labels were still associated with various task loads and the participants would still find a predictive model useful at times. The trimmed dataset contained 57.5% (347 of 603) "good" labels, a more balanced distribution. A useful classifying model should achieve a precision, if not accuracy, above this baseline number.

Table 2: Confusion matrices from tuning. We optimized for precision to reduce false positive predictions (top-left cell of each matrix).

(a) Default parameters. `scale_pos_weight` = 1
AUC ≈ 0.7516**. F1** ≈ 0.7299

		Predicted Label	
		Bad	Good
True Label	Bad	160	96
	Good	92	255

(b) Default parameters. `scale_pos_weight` = 0.4.
AUC ≈ 0.7537**. F1** ≈ 0.6520**.**

		Predicted Label	
		Bad	Good
True Label	Bad	205	51
	Good	154	193

(c) Default parameters. `scale_pos_weight` = 0.3.
AUC ≈ 0.7498**. F1** ≈ 0.6036**.**

		Predicted Label	
		Bad	Good
True Label	Bad	214	42
	Good	178	169

5.1 Classifier Training

The initial training of our classifier involved feature extraction, sample class balancing, algorithm selection, and time window sizing. We evaluated our choices based on the model performance in stratified 5-fold cross-validation (`StratifiedKFold` iterator from scikit-learn[4]; training sets were class-balanced with SMOTE[10]). We further explored our model by conservatively tuning it and investigating the importance of individual features. For reproducibility, all following steps had the `random_state` set to 1 if they contained any random seed.

5.1.1 Feature Definition. We extracted the features from the time-stamped sensor data streams discussed in 3.5, spanning $[t - t_w, t]$, where t_w is the size of the time window, and t is the moment of signal onset. We chose the moment of signal onset as the cut-off point because once the signal went off, participants' behavior could change before they provided a response, which introduces noise. The trade-off is that the data between the signal onset and the moment of response was unaccounted for.

We took the means and standard deviations of these data clips. For each clip, we calculated means, standard deviations, and "changes" (the difference between the mean of the data in $[t - \frac{t_w}{2}, t]$ and $[t - t_w, t - \frac{t_w}{2}]$ [61]). Additional features include the time since the last signal and the audio RMS levels of the video recordings. Table 3 is a complete list of training features.

5.1.2 Algorithm Selection. Two recent interruptibility studies, [54] and [61], used tree-based XGBoost [12] and Support Vector Machine (SVM) [28], respectively, to classify "good" and "bad" moments of interruption, yielding promising results. Additionally, we explored CatBoost [55], LightGBM [34], and the deep learning algorithm TabPFN [30], which are all efficient on local personal computers and require minimal training time. Initial tests (Table C.1) showed comparable performance across all candidates except SVM. We selected CatBoost for its strong support of categorical features compared to LightGBM and XGBoost and for its explainability and tunability compared to TabPFN. However, this choice is specific to our proof-of-concept development and does not rule out other algorithms for future applications.

5.1.3 Time Window Sizing. We trained the CatBoost classifier using default settings and compared the stratified 5-fold cross-validation performance using different lengths of time windows ranging from 5 to 20 seconds. We chose our time window t_w to be 15 seconds, as it yields the optimal performance given our data (Figure C.3 in Appendix B).

5.1.4 Conservative Tuning. Using a 15-second time window, the default CatBoost reached 70.15% accuracy, 74.18% precision, 0.7584 AUC, and 0.7406 F1 score in our 5-fold cross-validation (confusion matrix in Table 2a), already a good improvement over chance. Hyperopt [5] parameter optimization yielded marginal gains. Thus, we proceeded with default parameters, focusing on `scale_pos_weight` to reduce false-positive errors and increase precision. Lowering `scale_pos_weight` to 0.4 and 0.3 made the classifier more conservative, withholding notifications around half the time, reaching precisions of 79.13% and 80.30%, respectively (Tables 2b and 2c). Wu et al. [68] points out that in an automotive setting, a continuously running classifier will have a sufficient sample pool for many good notification opportunities. At `scale_pos_weight` = 0.1, the classifier became overly conservative, missing 255 of 347 "good" moments and dropping accuracy below 60% despite achieving >84% precision. Subsequent analyses used `scale_pos_weight` = 0.3.

[4] https://scikit-learn.org/

Table 3: List of features used in classifier training.

Data Source	Data Name	Components	Features	Note
Simulator	vehicle position (`veh_pos`), rotation (`veh_rot`), velocity (`veh_vel`), rotational velocity (`veh_rot_vel`),	x, y, z	mean, standard dev.	See Appendix C.1 for directions of the components
	wheel speed (`veh_wheel_speed`), steering angle (`veh_steering_angle`), audio output RMS (`disp_audio_rms`)	scalar		
	nearby vehicle count (`veh_nearby_vehicles`), nearby pedestrian count (`veh_nearby_peds`)	count		
Eye tracker	pupil center position (`pup_cent`), gaze direction (`gaze_dir`)	left, right × x, y, z	mean, standard dev., change	See Appendix C.1 for directions of the components
	3D gaze position (`3d_gaze_pos`)	x, y, z		
	pupil diameter (`pup_diam`)	left, right × scalar		
	Projected gaze position (`gaze_pos`)	x, y		Position of eye gaze projected onto the first-person-view video frame. See also Appendix C.1.
	FPV audio RMS (`gaze_audio_rms`)	scalar	mean, standard dev.	Audio level of the first-person-view video
Eye tracker MEMS	head accelerometer (`head_accel`), head gyroscope (`head_gyro`)	x, y, z	mean, standard dev., change	See Appendix C.1 for directions of the components
Tablet	tablet accelerometer (`tab_acc`)	x, y, z		
	touch (`tab_touch`)	count		
General information	scenario name (`scenario`), participant ID (`id`)	category		Used for grouping and feature addition experiments only
	time since the previous signal (`time_since_previous`)	scalar		

Table 4: Top 10 features from one 5-fold cross-validation, ranked by importance determined by CatBoost. All `random_state` variables were set to 1. The exact ranking varied slightly with the randomness during stratified sampling and tree split selection.

Feature Name	total_gain
head_accel_z_mean	3.361501
r_pup_cent_x_mean	3.055527
head_accel_y_mean	2.572980
head_gyro_x_std	2.527651
l_pup_cent_x_mean	2.114554
l_pup_cent_y_mean	1.356134
tab_touch_count	1.349825
tab_acc_z_std	1.291926
r_pup_diam_std	1.288371
r_pup_diam_mean_change	1.245845

5.1.5 Feature Importance. For each tune, we output the `feature_importances_` property, which shows "how much on average the prediction changes if the feature value changes[5]." Rankings of feature importance varied due to the randomness of cross-validation folds and tree selection during training. However, head pose and gaze direction consistently emerged as critical features. Table 4 is an example of the top 10 features for the model using `scale_pos_weight` = 0.3. Regardless of the random seed, head acceleration and rotation were consistently impactful; the tablet accelerometer readings and pupil diameter also influenced classification.

[5]https://catboost.ai/docs/en/concepts/fstr

5.1.6 Comparison between Train-Test Splitting Methods. In our initial model training, stratified sampling was used for cross-validation, meaning data from the same user and session often appeared in both the training and test sets (though never the same sample). This simulated a system familiar with a user's activities, making the reported results optimistic.

Thus, to assess performance in different contexts, we evaluated the CatBoost algorithm in three additional cases: 1) New users: Using `StratifiedGroupKFold`, data from a single participant appeared in either the training or test set, but not both. 2) Known users, new activities: Using `LeaveOneGroupOut`, all data from one specific scenario (e.g., Game, Video, Reading) was assigned to the test set. 3) Known users, familiar activities: Using `StratifiedGroupKFold`, data was grouped by participant ID and scenario, ensuring that each session was exclusive to either the training or test set.

Because there were only four scenarios, the cross-validation for Case 2) could only be 4-fold. For a fairer comparison, we performed 4-fold cross-validations in all cases, including the original train-test split. Results (Table 5) showed worsening performance—accuracy close to chance despite > 70% precision. However, this does not negate the value of our initial findings, as discussed further in the Discussion section.

5.1.7 Ablation Studies. To illustrate the effects of head pose measurements and gaze measurements, we ran a "rough" ablation study for the related features. Removing gaze- and pupil-related features noticeably reduced model performance (75.87% precision, 60.86% accuracy, AUC= 0.6913, and F1= 0.5797). Removing the all head and

Table 5: Train-test split method comparison (4-fold cross validation). All CV runs used the 15s time window and the tuned parameters `scale_pos_weight` = 0.3, optimizing for precision. All steps had `random_state` = 1 if any random seed was involved.

Split Method	Explanation	Precision	Accuracy	AUC	F1 Score
Stratified Group *k*-Fold w/ `id`	The system has never seen this user.	71.55%	57.24%	0.6089	0.5488
Leave One Group Out w/ `scenario`	The system knows this user, but has never seen this activity.	72.38%	60.57%	0.6762	0.5687
Stratified Group *k*-Fold w/ `id` + `scenario`	The system is familiar with the activity and the user but hasn't seen them together before.	74.05%	58.51%	0.6705	0.5542
Stratified *k*-Fold	The system has seen this user doing this activity.	79.47%	62.52%	0.7485	0.5910

eye tracking measurements resulted in predictions no better than chance (See Figures C.5a and C.5b in Appendix B). Next, a more detailed ablation study evaluated the individual contribution of 20 behavioral measurement groups (excluding `id` and `scenario` from Table 3). Most removals caused minor performance changes (AUC ±0.01), except pupil center position and gaze direction, which reduced AUC by -0.0277 and -0.0124, respectively, further suggesting the importance of eye and gaze measurements.

5.1.8 Inclusion of User and Activity Features. We also explored the effect of including participant identity (`id`) and activity (`scenario`) as categorical features, simulating modern facial recognition [17] and activity identification [7, 21, 42] technologies. Adding `id` and `scenario` yielded 62.68% accuracy, 79.91% precision, and 0.7489 AUC in stratified 5-fold cross-validation—similar to the initial model. Notably, `id` showed strong feature importance (3.05, second to `head_accel_z_mean`). This indicates that the model could heavily rely on the participant ID, although it could also perform similarly well without such information. Nevertheless, `scenario` had moderate importance (1.15), reflecting some potential predictive value of activity detection.

5.1.9 Inclusion of Always-Available Participants. We noticed that including data from participants who always answered "good," despite NDRT, could yield more optimistic, but less generalizable results. To demonstrate, we trained a model using all available data. We included participant ID (`id`) and activity (`scenario`), set `scale_pos_weight` to 0.3. The model reached 71.43% accuracy (c.f. 63.4% chance), 88.09% precision, and 0.8316 AUC in stratified 5-fold cross-validation. The model became over-reliant on `id`, with a very high importance of 12.36 (c.f. 3.05 above). Readers may refer to Figure C.6 in Appendix B for the corresponding ROC curve.

5.2 Correlation Studies

Another approach to evaluate feature importance is by analyzing statistical correlations between variables. Tables 6a and 6b, as well as Tables C.2a and C.2b in Appendix B list the features most strongly correlated with "good/bad" labels and with signal detection, response, and decision times. All features were derived from the 15-second time window preceding each audio signal. Though correlations were not strong overall, many were statistically significant. Consistent with classifier training results, head poses, gaze direction, and tablet position correlated with participants' **availability** to receive non-emergency notifications, while vehicle status had less influence.

Table 6: Correlations (a) between features and "good" labels from the participants and (b) between features and time needed for the participants to detect the audio signal. Only the top 20 features are shown.

(a) Point-biserial correlations between features and "good" labels. *: $p < 0.05$. **: $p < 0.01$.

Feature Name	Corr. w/"good" labels
`head_gyro_x_std`	0.241**
`head_accel_z_mean`	-0.234**
`head_accel_z_std`	0.234**
`head_gyro_y_std`	0.217**
`r_gaze_dir_y_std`	0.204**
`gaze_pos_y_std`	0.199**
`r_pup_cent_x_std`	0.191**
`head_accel_y_std`	0.172**
`l_pup_diam_mean`	-0.167**
`r_gaze_dir_x_std`	0.165**
`gaze_pos_x_std`	0.163**
`head_gyro_z_std`	0.160**
`r_pup_diam_mean`	-0.158**
`l_gaze_dir_y_std`	0.152**
`tab_touch_count`	-0.148**
`l_pup_cent_z_mean`	0.134**
`r_pup_cent_z_mean`	0.133**
`l_pup_cent_y_mean`	-0.129**
`l_pup_cent_x_std`	0.127**
`r_pup_cent_y_mean`	-0.125**

(b) Pearson correlations between features and audio signal detection time measurements. *: $p < 0.05$. **: $p < 0.01$.

Feature Name	Corr. w/ $t_{\text{detection}}$
`disp_audio_rms_mean`	0.336**
`gaze_audio_rms_mean`	0.279**
`tab_touch_count`	-0.244**
`veh_vel_x_mean`	0.195**
`veh_rot_y_mean`	-0.167**
`tab_acc_z_mean`	0.167**
`tab_acc_y_mean`	-0.166**
`r_pup_cent_y_std`	0.161**
`tab_acc_z_std`	-0.160**
`veh_nearby_vehicles_count_mean`	0.160**
`r_gaze_dir_z_mean_change`	0.152**
`r_pup_cent_y_mean`	0.151**
`gaze_audio_rms_std`	0.148**
`r_pup_cent_x_mean`	0.145**
`disp_audio_rms_std`	0.145**
`l_pup_cent_y_std`	0.144**
`gaze_pos_y_mean`	0.135**
`l_pup_cent_z_mean`	0.134**
`veh_pos_y_mean`	0.134**
`3d_gaze_pos_x_mean`	0.134**

The correlation between the features and signal detection time shows that the participants' **ability** to detect and respond to the sound signal depended on various factors:

(1) *Ambient noises*: Detection time increased with loud external sounds, e.g., trains or motorcycles.
(2) *Tablet touches*: Watching videos led to fewer tablet touches and occupied audio channels, delaying detection and response, as confirmed in Section 4.2.
(3) *Vehicle status*: Possibly related to our course design, the drive's main segment involved loud highway noise, more nearby vehicles, and frequent right turns. In these cases, participants were absorbed into their NDRTs, causing a longer detection time. Only towards the end of the drive did most participants disengage from their tasks.
(4) *Gaze direction*: Slower detection occurred when participants focused on lower, closer objects, reflected by pupil position (`pup_cent`) and gaze direction (`gaze_dir`). The eye tracker also tended to slide away from the face at such moments, increasing `pup_cent_z` measurements.

Head pose did not significantly affect signal detection times, contrasting its role in predicting the subjective "good/bad" labels. The "change" features showed limited influence on either group of dependent variables.

The features affecting participants' response time were similar to those affecting detection time in that the former contains the latter. Regarding decision time (response time minus detection time), participants in the Game scenario made quicker decisions with fewer actions due to high engagement and time pressure (Section 4.2). Thus, tablet touches and pupil diameters showed stronger correlations with decision time. Reaction time correlations were mostly very weak. No statistically significant correlation was found between "good" labels and detection times.

6 Discussion

Our study allowed participants to experience a future automotive scenario and elaborate on how they assessed their availability for non-emergency notifications, answering RQ.1. Our trained CatBoost classifier, albeit basic compared to deep learning models, showed preliminary effectiveness in predicting notification timing, supporting our hypothesis for RQ.2. Here we discuss key takeaways in relation to past work and implications for future design.

6.1 Comparison to Similar Studies

In this study, we adapted the interruptibility study format [23, 31, 54] to a simulated AV environment [8, 9, 22] to reveal the participants' different behavior patterns when receiving in-vehicle notifications. Our results highlight similarities and differences between autonomous and manual driving notification systems. In both, drivers' actions (e.g., tablet touches or hands on the wheel [36]) reflect their tasks and engagement, aiding interruptibility prediction. However, in autonomous driving, road factors only influence signal noticeability, not interruptibility as in manual driving.

Prediction in autonomous vehicles required a longer time window (15 seconds, compared to ≈ 2 seconds [35] and 6 seconds [68] in manual driving). This aligns with interruptibility studies in office environments, such as Hudson et al. [31] (≈ 15s) and Züger and Fritz [75] (> 10s). However, unlike manual driving or office settings, AVs may enable more diverse postures, including extreme seating positions that occasionally occlude participants' faces (Figures B.2b, B.4, and B.5 in Appendix B.4). As reclined, ergonomically beneficial postures become more common in AVs [6, 51], robust and flexible sensor configurations will be necessary.

6.2 Predicting Notification Timing in the Real World

We reflect here on the feasibility of extracting features to predict appropriate notification timing in the real world.

6.2.1 Gaze and Posture Estimation. In self-driving car UX research, gaze and posture estimation have been implemented primarily for safety and quality concerns of takeover [16, 29, 70]. These works showcased the potential of sensing and prediction technologies in future vehicles. Our study shows that gaze and head pose measurements also can, to some extent, indicate when occupants engaged in NDRTs may prefer receiving non-emergency notifications.

Even without specialized eye-tracking devices, our concept remains feasible in vehicles equipped with occupant sensing systems like vision sensors and seat sensors. Models such as GazeDPTR[13], OpenFace[2], and GazeML[52] effectively estimate human gaze under dynamic conditions. Recent advancements, such as Tambwekar et al. [64], demonstrate computer vision-based occupant posture estimation, while studies such as [38] and [73] employed seat-mounted sensors to assess posture in AVs.

6.2.2 User Activity on Smart Devices. Our results show that device usage, including tablet content, touches, and orientation, also helps with non-emergency notification timing. Studies such as [54] and [71] leveraged everyday mobile device activity to train models predicting the opportune moments for notifications. As drivers interact with connected devices such as smartphones and tablets, these techniques can be adapted to automotive environments to improve notification timing predictions.

6.2.3 Cognitive Load Estimation. Evidence in our study suggests that real-time cognitive load indicators (e.g. pupil diameter and potentially hunched posture [32]) may be predictors of when to send notifications to drivers. Using data from wearable psychophysiological sensors, Züger and Fritz [75] reached a similar conclusion in a desk job interruptibility study. We speculate that in our case, having more reliable yet non-intrusive cognitive load estimators will benefit the timing predictions. Computer vision [26] or infrared sensing [1] can be good contactless options in automotive cabins.

6.2.4 New Users and Activities. Our proof-of-concept model showed performance degradation when faced with new users and/or new activities (Table 5). We argue that should predictive notification timing become a reality, the model will include a large amount of diverse data and continue to learn from new users incrementally. Incremental lifelong learning has already been used to improve driving behavior in autonomous cars, which constantly acquire new data for safety-critical situations [74]. The smart agent timing in-vehicle notifications may initially collect new users' preferences and occasionally invite user feedback after pushing notifications.

Thus, designers may consider different ways to collect user feedback and thereby use the information as additional data labels.

6.3 Implications for Design

Results from the classifier training, correlation studies, and qualitative observations help us address RQ.3 on design considerations. Due to more diverse user behaviors in AVs, delivering notifications with proper timing can become challenging. We propose three aspects to be considered for well-timed notifications and interactions:

6.3.1 Design for Individuals. For manual driving, "good" moments typically align with simpler traffic conditions and steady vehicle movements [61]. In autonomous vehicles, however, our data reveal personal nuances in interpreting "good" moments. Certain individuals may always be mentally available for the *make-aware* or *change-blind* notifications, while others have sporadic availability. Despite subjective differences, commonalities like head pose, gaze direction, and tablet activity correlate with "good" moments, enabling promising prediction precision in our classifier. Future UX designs can leverage these shared patterns as sensing technologies advance. **Providing drivers with options such as "notification always on" and "intelligent notification" may be an effective starting point.**

6.3.2 Design for Tasks. NDRTs influence both a driver's **ability** to detect a signal and their **availability** for notifications. Even with the same device, tasks induce varying cognitive demands, reflected in gaze and posture changes, which affect notification timing. Understanding the effects of different NDRTs on a driver's behavior can contribute to better-timed virtual content delivery in the vehicle. **In "intelligent notification" mode, the smart agent can estimate engagement, cognitive load, activity, and task progress. It can utilize multi-modal sensors and cues to encompass different types of NDRTs. It can also refine its behavior through occasional user feedback, particularly when encountering unfamiliar tasks.**

6.3.3 Design for Contexts. Our study showed that environmental factors, such as road noises, correlate with a driver's ability to detect audio notification cues (Section 5.2). Usage time and familiarity also had an impact (Section 4.2). Meanwhile, implicit factors can affect drivers' availability. Survey responses in Section 4.3.2 suggest that drivers may exhibit different preferences due to their mood (mental state) or events before, during, or after the drive (causing fatigue or stress). **Future "intelligent notification" modes can explore space and time beyond the vehicle and its current journey.**

7 Limitations and Future Work

This study faced limitations due to its static desktop driving simulator lacking locomotion feedback. Future research could explore how vehicle motion influences attention and notification timing. To ensure generalizability, we did not present any specific notification content to provide additional context. Future work may explore the interaction effect of notification content. While a four-day longitudinal study was conducted, longer-term interactions with AVs and NDRTs like eating or exercising [53] may require further investigation to reveal other important features. Finally, our model still has room for improvement. We selected a fast, explainable algorithm as a proof-of-concept baseline. We also chose simple binary labels instead of scale ratings to represent timing appropriateness. Future AVs leveraging deep learning may benefit from raw time-series inputs and training datasets with ordinal- or numerical-scale labels, enabling finer-grained prioritization of non-emergency notifications.

Studying virtual content timing in the context of autonomous driving is still new. For our next step, we hope to develop a more immersive simulation environment with adaptive notification timing and content. Incremental learning [49] may improve personalization, and computer vision-based techniques like OpenFace 2.0 [2] could replace eye tracking to enhance sensing capabilities.

8 Conclusion

Future autonomous vehicles (AVs) will allow drivers and passengers to engage in various non-driving-related tasks (NDRTs). In this project, we designed an interruptibility study on a desktop autonomous driving simulator and investigated the possibility of scheduling non-emergency notifications more appropriately for drivers doing NDRTs. We identified four driver categories based on availability for notifications: always available, prioritizing NDRTs, task content dependent, and mental state dependent. Using behavioral and vehicular data with participant-supplied labels, we trained a proof-of-concept machine learning classifier that achieved effective precision in predicting notification timing. Key features for prediction included head pose, eye gaze, and mobile device usage, measurable by intelligent sensors in future AVs. Together with qualitative observations and survey responses, these findings provide insights into the Individual, Task, and Context dimensions of content delivery timing in autonomous vehicles.

Acknowledgments

This work was generously supported by funding from the Honda Research Institute, USA, with special thanks to Erin Clepper, Joan Smith, and Duane Detwiler. The authors thank David Lindlbauer and Yi Fei Cheng at Carnegie Mellon University for the initial project concept. Henny Admoni and Laura Dabbish at CMU kindly provided the eye-tracking equipment. The authors also received advice from Jorge Ortiz at Rutgers University on human behavior modeling.

The authors thank Joshua Winfrey, a graduate of the Master of Human-Computer Interaction program at CMU, for his invaluable assistance in pilot testing sessions and refining the user study protocols. Additional thanks to Rainer Hilmer (aka `Cyron43`), a game modification developer who provided the original code base enabling the simulated autonomous driving experience.

References

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 33:1–33:20. doi:10.1145/3130898

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, Xi'an, China, 59–66. doi:10.1109/FG.2018.00019

[3] Siddhartha Banerjee, Andrew Silva, and Sonia Chernova. 2018. Robot Classification of Human Interruptibility and a Study of Its Effects. *J. Hum.-Robot Interact.* 7, 2 (Oct. 2018), 14:1–14:35. doi:10.1145/3277902

[4] Martin R. K. Baumann, Diana Rösler, and Josef F. Krems. 2007. Situation Awareness and Secondary Task Performance While Driving. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris (Ed.). Springer, Berlin, Heidelberg, 256–263. doi:10.1007/978-3-540-73331-7_27

[5] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 115–123. https://proceedings.mlr.press/v28/bergstra13.html

[6] Dominique Bohrmann and Klaus Bengler. 2020. Reclined Posture for Enabling Autonomous Driving. In *Human Systems Engineering and Design II*, Tareq Ahram, Waldemar Karwowski, Stefan Pickl, and Redha Taiar (Eds.). Springer International Publishing, Cham, 169–175. doi:10.1007/978-3-030-27928-8_26

[7] Christian Braunagel, Enkelejda Kasneci, Wolfgang Stolzmann, and Wolfgang Rosenstiel. 2015. Driver-Activity Recognition in the Context of Conditionally Autonomous Driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, Gran Canaria, Spain, 1652–1657. doi:10.1109/ITSC.2015.268

[8] Gary Burnett and Davide Salanitri. 2019. *How will drivers interact with vehicles of the future?* Technical Report. RAC Foundation. https://www.racfoundation.org/wp-content/uploads/Automated_Driver_Simulator_Report_July_2019.pdf

[9] Marine Capallera, Quentin Meteier, Emmanuel De Salis, Marino Widmer, Leonardo Angelini, Stefano Carrino, Andreas Sonderegger, Omar Abou Khaled, and Elena Mugellini. 2023. A Contextual Multimodal System for Increasing Situation Awareness and Takeover Quality in Conditionally Automated Driving. *IEEE Access* 11 (2023), 5746–5771. doi:10.1109/ACCESS.2023.3236814

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357.

[11] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th international conference on Intelligent user interfaces (IUI '11)*. Association for Computing Machinery, New York, NY, USA, 315–318. doi:10.1145/1943403.1943454

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785

[13] Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. 2024. What Do You See in Vehicle? Comprehensive Vision Solution for In-Vehicle Gaze Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, 1556–1565. doi:10.1109/CVPR52733.2024.00154

[14] Damee Choi, Toshihisa Sato, Takafumi Ando, Takashi Abe, Motoyuki Akamatsu, and Satoshi Kitazaki. 2020. Effects of cognitive and visual loads on driving performance after take-over request (TOR) in automated driving. *Applied Ergonomics* 85 (May 2020), 103074. doi:10.1016/j.apergo.2020.103074

[15] Vera Demberg, Asad Sayeed, Angela Mahr, and Christian Müller. 2013. Measuring linguistically-induced cognitive load during driving using the ConTRe task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*. Association for Computing Machinery, New York, NY, USA, 176–183. doi:10.1145/2516540.2516546

[16] Nachiket Deo and Mohan M. Trivedi. 2020. Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness. *IEEE Transactions on Intelligent Vehicles* 5, 1 (March 2020), 41–52. doi:10.1109/TIV.2019.2955364

[17] Ekberjan Derman and Albert Ali Salah. 2018. Continuous Real-Time Vehicle Driver Authentication Using Convolutional Neural Network Based Face Recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, Xi'an, China, 577–584. doi:10.1109/FG.2018.00092

[18] Henrik Detjen, Bastian Pfleging, and Stefan Schneegass. 2020. A Wizard of Oz Field Study to Understand Non-Driving-Related Activities, Trust, and Acceptance of Automated Vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 19–29. doi:10.1145/3409120.3410662

[19] Mica R. Endsley and Daniel J. Garland (Eds.). 2000. *Situation awareness: analysis and measurement*. Lawrence Erlbaum Associates, Mahwah, NJ.

[20] Mica R. Endsley and Debra G. Jones. 2012. *Designing for situation awareness: an approach to user-centered design* (second edition ed.). CRC Press, Taylor & Francis Group, Boca Raton London New York.

[21] Saad Ezzini, Ismail Berrada, and Mounir Ghogho. 2018. Who is behind the wheel? Driver identification and fingerprinting. *Journal of Big Data* 5, 1 (Feb. 2018), 9. doi:10.1186/s40537-018-0118-7

[22] Sarah Faltaous, Chris Schönherr, Henrik Detjen, and Stefan Schneegass. 2019. Exploring proprioceptive take-over requests for highly automated vehicles. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3365610.3365644

[23] Joel Fischer. 2011. *Understanding receptivity to interruptions in mobile human-computer interaction*. Thesis (University of Nottingham only). University of Nottingham. https://eprints.nottingham.ac.uk/12499/

[24] Daniel Fitousi and Michael J. Wenger. 2011. Processing capacity under perceptual and cognitive load: A closer look at load theory. *Journal of Experimental Psychology: Human Perception and Performance* 37, 3 (2011), 781–798. doi:10.1037/a0020675

[25] International Organization for Standardization. 2019. Road vehicles — Ergonomic aspects of transport information and control systems — Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems (ISO/TS 14198:2019). https://www.iso.org/standard/71509.html

[26] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive Load Estimation in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3173574.3174226

[27] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. doi:10.1177/154193120605000909

[28] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13, 4 (July 1998), 18–28. doi:10.1109/5254.708428

[29] Ala'aldin Hijaz, Wing-Yue Geoffrey Louie, and Iyad Mansour. 2019. Towards a Driver Monitoring System for Estimating Driver Situational Awareness. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New Delhi, India, 1–6. doi:10.1109/RO-MAN46459.2019.8956378

[30] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. doi:10.48550/arXiv.2207.01848

[31] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting human interruptibility with sensors: a Wizard of Oz feasibility study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 257–264. doi:10.1145/642611.642657

[32] Go Igarashi, Chieko Karashima, and Minoru Hoshiyama. 2016. Effect of Cognitive Load on Seating Posture in Children. *Occupational Therapy International* 23, 1 (2016), 48–56. doi:10.1002/oti.1405

[33] International Organization for Standardization. 2016. Road vehicles — Transport information and control systems — Detection-response task (DRT) for assessing attentional effects of cognitive load in driving (ISO 17488:2016). https://www.iso.org/standard/59887.html

[34] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, CA, 3146–3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[35] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (Dec. 2018), 175:1–175:28. doi:10.1145/3287053

[36] SeungJun Kim, Jaemin Chun, and Anind K. Dey. 2015. Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 487–496. doi:10.1145/2702123.2702409

[37] Sang Min Ko and Yong Gu Ji. 2018. How we can measure the non-driving-task engagement in automated driving: Comparing flow experience and workload. *Applied Ergonomics* 67 (Feb. 2018), 237–245. doi:10.1016/j.apergo.2017.10.009

[38] Rahul Prasanna Kumar, David Melcher, Pietro Buttolo, and Yunyi Jia. 2023. Tracking Occupant Activities in Autonomous Vehicles Using Capacitive Sensing. *IEEE Transactions on Intelligent Transportation Systems* 24, 7 (July 2023), 6800–6819. doi:10.1109/TITS.2023.3266000

[39] Antonio Luque-Casado, José C. Perales, David Cárdenas, and Daniel Sanabria. 2016. Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology* 113 (Jan. 2016), 83–90. doi:10.1016/j.biopsycho.2015.11.013

[40] Andreas Löcken, Shadan Sadeghian Borojeni, Heiko Müller, Thomas M. Gable, Stefano Triberti, Cyriel Diels, Christiane Glatz, Ignacio Alvarez, Lewis Chuang, and Susanne Boll. 2017. Towards Adaptive Ambient In-Vehicle Displays and Interactions: Insights and Design Guidelines from the 2015 AutomotiveUI Dedicated Workshop. In *Automotive User Interfaces: Creating Interactive Experiences in the Car*, Gerrit Meixner and Christian Müller (Eds.). Springer International Publishing, Cham, 325–348. doi:10.1007/978-3-319-49448-7_12

[41] Dawn C. Marshall, John D. Lee, and P. Albert Austria. 2007. Alerts for In-Vehicle Information Systems: Annoyance, Urgency, and Appropriateness. *Human Factors* 49, 1 (Feb. 2007), 145–157. doi:10.1518/001872007779598145

[42] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon ReiB, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 2801–2810. doi:10.1109/ICCV.2019.00289

[43] Tara Matthews, Anind K. Dey, Jennifer Mankoff, Scott Carter, and Tye Rattenbury. 2004. A toolkit for managing user attention in peripheral displays. In *Proceedings of the 17th annual ACM symposium on User interface software and technology (UIST '04)*. Association for Computing Machinery, New York, NY, USA, 247–256. doi:10.1145/1029632.1029676

[44] Vadim Melnicuk, Simon Thompson, Paul Jennings, and Stewart Birrell. 2021. Effect of cognitive load on drivers' State and task performance during automated driving: Introducing a novel method for determining stabilisation time following take-over of control. *Accident Analysis & Prevention* 151 (March 2021), 105967. doi:10.1016/j.aap.2020.105967

[45] David Miller, Annabel Sun, Mishel Johns, Hillary Ive, David Sirkin, Sudipto Aich, and Wendy Ju. 2015. Distraction Becomes Engagement in Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (Sept. 2015), 1676–1680. doi:10.1177/1541931215591362

[46] Andreas Lars Müller, Natacha Fernandes-Estrela, Ruben Hetfleisch, Lukas Zecha, and Bettina Abendroth. 2021. Effects of non-driving related tasks on mental workload and take-over times during conditional automated driving. *European Transport Research Review* 13, 1 (Feb. 2021), 16. doi:10.1186/s12544-021-00475-5

[47] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. 2018. A Review of Non-driving-related Tasks Used in Studies on Automated Driving. In *Advances in Human Aspects of Transportation*, Neville A Stanton (Ed.). Springer International Publishing, Cham, 525–537. doi:10.1007/978-3-319-60441-1_52

[48] Frederik Naujoks, Andrea Kiesel, and Alexandra Neukum. 2016. Cooperative warning systems: The impact of false and unnecessary alarms on drivers' compliance. *Accident Analysis & Prevention* 97 (Dec. 2016), 162–175. doi:10.1016/j.aap.2016.09.009

[49] Hyungik Oh, Laleh Jalali, and Ramesh Jain. 2015. An intelligent notification system using context from real-time personal activity monitoring. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Turin, Italy, 1–6. doi:10.1109/ICME.2015.7177508

[50] Erfan Pakdamanian, Nauder Namaky, Shili Sheng, Inki Kim, James Arthur Coan, and Lu Feng. 2020. Toward Minimum Startle After Take-Over Request: A Preliminary Study of Physiological Data. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 27–29. doi:10.1145/3409251.3411715

[51] Sibashis Parida, Sai Mallavarapu, Sylvester Abanteriba, Matthias Franz, and Wolfgang Gruener. 2019. Seating Postures for Autonomous Driving Secondary Activities. In *Innovation in Medicine and Healthcare Systems, and Multimedia*, Yen-Wei Chen, Alfred Zimmermann, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer, Singapore, 423–434. doi:10.1007/978-981-13-8566-7_39

[52] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 2018. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3204493.3204545

[53] Bastian Pfleging, Maurice Rang, and Nora Broy. 2016. Investigating user needs for non-driving-related activities during automated driving. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia (MUM '16)*. Association for Computing Machinery, New York, NY, USA, 91–99. doi:10.1145/3012709.3012735

[54] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 91:1–91:25. doi:10.1145/3130956

[55] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., Montréal, Canada, 6638–6648. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

[56] Jonas Radlmayr, Christian Gold, Lutz Lorenz, Mehdi Farid, and Klaus Bengler. 2014. How Traffic Situations and Non-Driving Related Tasks Affect the Take-Over Quality in Highly Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (Sept. 2014), 2063–2067. doi:10.1177/1541931214581434

[57] Andreas Riener, Myounghoon Jeon, Ignacio Alvarez, and Anna K. Frison. 2017. Driver in the Loop: Best Practices in Automotive Sensing and Feedback Mechanisms. In *Automotive User Interfaces: Creating Interactive Experiences in the Car*, Gerrit Meixner and Christian Müller (Eds.). Springer International Publishing, Cham, 295–323. doi:10.1007/978-3-319-49448-7_11

[58] Noelia Rivera-Garrido, M. P. Ramos-Sosa, Michela Accerenzi, and Pablo Brañas-Garza. 2022. Continuous and binary sets of responses differ in the field. *Scientific Reports* 12, 1 (Aug. 2022), 14376. doi:10.1038/s41598-022-17907-4

[59] SAE International. 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016_202104). doi:10.4271/J3016_202104

[60] Matti Schwalk, Niko Kalogerakis, and Thomas Maier. 2015. Driver Support by a Vibrotactile Seat Matrix – Recognition, Adequacy and Workload of Tactile Patterns in Take-over Scenarios During Automated Driving. *Procedia Manufacturing* 3 (Jan. 2015), 2466–2473. doi:10.1016/j.promfg.2015.07.507

[61] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300867

[62] David Sirkin, Nikolas Martelaro, Mishel Johns, and Wendy Ju. 2017. Toward Measurement of Situation Awareness in Autonomous Vehicles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 405–415. doi:10.1145/3025453.3025822

[63] Kristina Stojmenova, Grega Jakus, and Jaka Sodnik. 2017. Sensitivity evaluation of the visual, tactile, and auditory detection response task method while driving. *Traffic Injury Prevention* 18, 4 (May 2017), 431–436. doi:10.1080/15389588.2016.1214868

[64] Anuj Tambwekar, Byoung-Keon D. Park, Arpan Kusari, and Wenbo Sun. 2024. Three-Dimensional Posture Estimation of Vehicle Occupants Using Depth and Infrared Images. *Sensors* 24, 17 (Jan. 2024), 5530. doi:10.3390/s24175530

[65] Simon Trask, Madeline Stewart, Thomas Kerwin, Shawn Midlam-Mohler, Simon Trask, Madeline Stewart, Thomas Kerwin, and Shawn Midlam-Mohler. 2019. *Effectiveness of Warning Signals in Semi-Autonomous Vehicles*. Technical Report. SAE International, Warrendale, PA, USA. doi:10.4271/2019-01-1013

[66] Michael A. Vidulich and Pamela S. Tsang. 2012. Mental Workload and Situation Awareness. In *Handbook of Human Factors and Ergonomics*. John Wiley & Sons, Ltd, Hoboken, NJ, 243–273. doi:10.1002/9781118131350.ch8

[67] Bob G. Witmer and Michael J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (June 1998), 225–240. doi:10.1162/105474698565686

[68] Tong Wu, Nikolas Martelaro, Simon Stent, Jorge Ortiz, and Wendy Ju. 2021. Learning When Agents Can Talk to Drivers Using the INAGT Dataset and Multisensor Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (Sept. 2021), 133:1–133:28. doi:10.1145/3478125

[69] Jing Yang, Nade Liang, Brandon J. Pitts, Kwaku O. Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Denny Yu. 2023. Multimodal Sensing and Computational Intelligence for Situation Awareness Classification in Autonomous Driving. *IEEE Transactions on Human-Machine Systems* 53, 2 (April 2023), 270–281. doi:10.1109/THMS.2023.3234429

[70] Lichao Yang, Kuo Dong, Arkadiusz Jan Dmitruk, James Brighton, and Yifan Zhao. 2020. A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring. *IEEE Transactions on Intelligent Transportation Systems* 21, 10 (Oct. 2020), 4318–4327. doi:10.1109/TITS.2019.2939676

[71] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5346–5360. doi:10.1145/3025453.3025946

[72] Kathrin Zeeb, Axel Buchner, and Michael Schrauf. 2016. Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident Analysis & Prevention* 92 (July 2016), 230–239. doi:10.1016/j.aap.2016.04.002

[73] Mingming Zhao, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2021. Driver posture monitoring in highly automated vehicles using pressure measurement. *Traffic Injury Prevention* 22, 4 (April 2021), 278–283. doi:10.1080/15389588.2021.1892087

[74] Dekang Zhu, Qianyi Bu, Zhongpan Zhu, Yujie Zhang, and Zhipeng Wang. 2024. Advancing autonomy through lifelong learning: a survey of autonomous intelligent systems. *Frontiers in Neurorobotics* 18 (April 2024). doi:10.3389/fnbot.2024.1385778

[75] Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2981–2990. doi:10.1145/2702123.2702593

A Supplemental Information on Data Collection

A.1 Driving Course Design

Figure A.1 shows the driving course used in our study. It was designed to simulate a commute that goes through various environments. The commute started from a hotel in an urban area near the west coast of the map (location a, Banner Hotel & Spa, Del Perro) and went along the highway (location b, Grand Ocean Highway, Lago Zancudo) on the west and north coasts, which also passes through a coastal town (location c, Paleto Bay). Then it continued through the highway (location d, Grand Ocean Highway, Mount Gordo) and turned to the rural desert roads (location e, Joshua Road, Sandy Shores) in the northeastern part of the map. The course extended westwards through the area and then southwards through a valley (location f, Tongva Drive, Tongva Valley) and the suburbs (location g, Richman). It finally took a winding road that led to a corporate campus (location h, Kortz Center, Pacific Bluffs) back in the west part of the map. Overall, the drive formed a clockwise loop and lasted approximately 25 minutes.

For the situational awareness task discussed in 3.4, the random objects would appear close to locations b, c, d, e, f, and g, which are almost evenly distributed along the route in terms of the travel time.

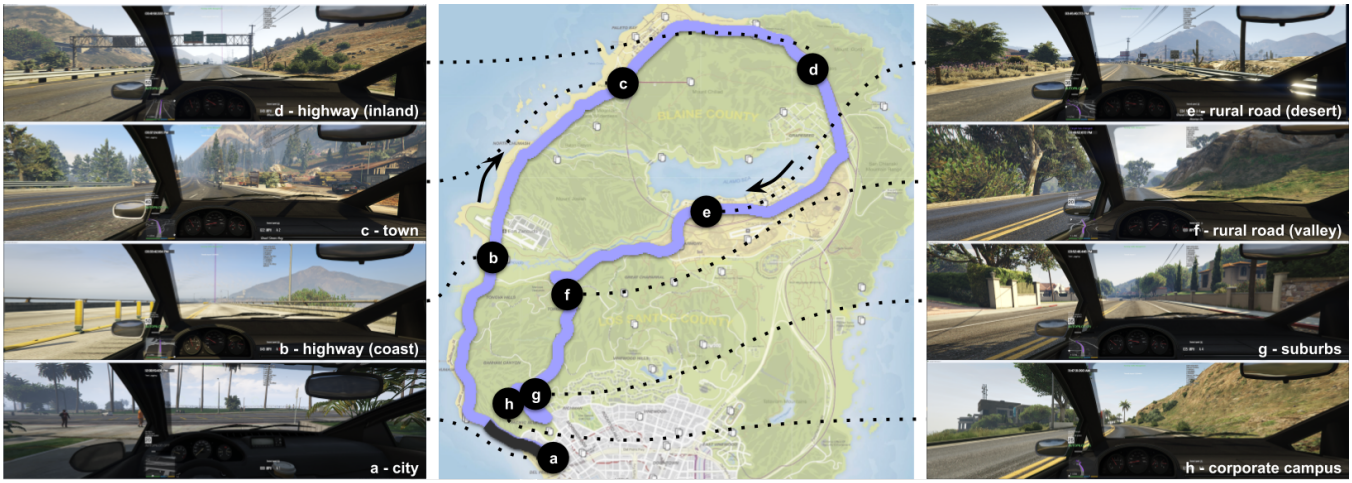


Figure A.1: Simulated commute course, which covered city, highway, town, rural, and suburban driving. The dark-colored portion of the route connected the city area to the highway and was manual-driving only; participants were prompted to turn on autopilot afterward.

A.2 Illustration of Data Labeling Procedure

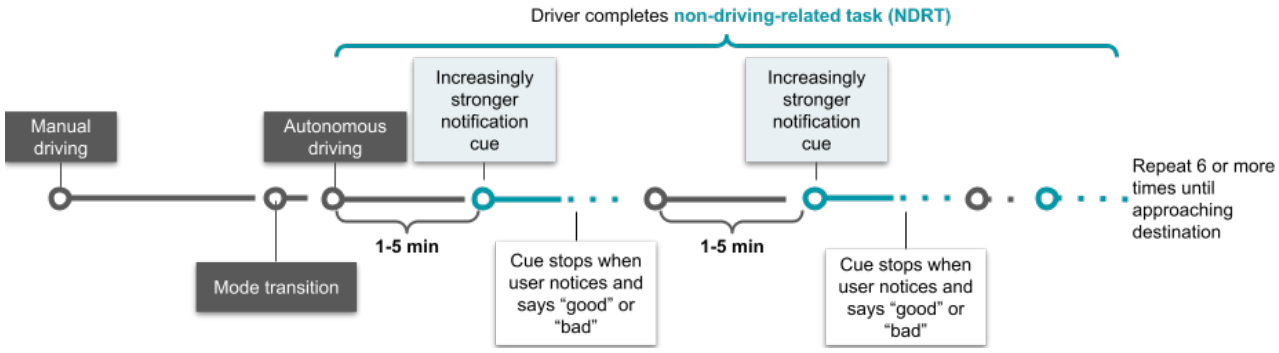


Figure A.2: Data labeling procedure.

A.3 Instruction Script

Before the start of each drive, the researcher restated the purpose of the study and the data labeling approach using the following script:

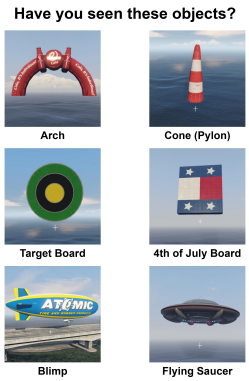
> "In the future, most drivers will be doing tasks unrelated to driving when the car is in autonomous driving mode. There will be moments when the car wants to notify you of certain things you might want to be aware of, but don't have to respond to immediately. This may include traffic conditions, charging station locations, non-emergency takeover requests, etc. We are investigating when to push such non-emergency notifications to drivers in autonomous cars. We don't want to startle the driver or distract them from doing their own thing too much."
>
> "In each session of this study, we would like to simulate a commuting scenario so you can help us with data collection. You will try out our driving simulator, switch to autopilot mode, and do some tasks while responding to a sound signal. The sound signal goes off at random times, and you will let us know if those are good moments to send you notifications."

The researcher then explicitly explained the definitions of a "good" and a "bad" moment to the participant:

> "You may say 'good/yes' if you don't find a notification at the moment distracting or annoying—especially regarding your own or assigned task; say 'bad/no' if otherwise, or if you don't want to be disturbed. In any case, keep your response in one word. Please keep in mind that your judgment is irrelevant to the sound quality or the road condition."

A.4 Situational Awareness Task

We selected a list of six objects (inflatable arch, inflatable pylon, floating round target board, floating square board, blimp, flying saucer) to be placed at six random locations on the route (roughly evenly distributed). For each session, we randomly inserted three of the objects into the course at three of the six locations. The list of objects was shown to each participant before the drive; a copy of the list was also placed on the wall behind the computer screen for later reference. Participants were instructed to verbally confirm that they saw any of the objects during the drive or to indicate what they saw in the post-session survey. This task was partially inspired by the Daze platform that occasionally asked AV occupants if they noticed an event or an object on the road [62]. However, since this is not the main focus of our study, we limited the number of objects shown to the participants. Figure A.3 shows the list of random objects and an example of one present on the road.



(a) The list of random objects. A participant would have access to the list for reference before and throughout the drive.

(b) One of the objects, an inflatable arch, appeared on the road. The participant failed to detect it due to the NDRT.

Figure A.3: Situational awareness task.

A.5 Timing Measurements

Figure A.4 illustrates the four timing measurements collected from the video data: detection time, response time, reaction time, and decision time.

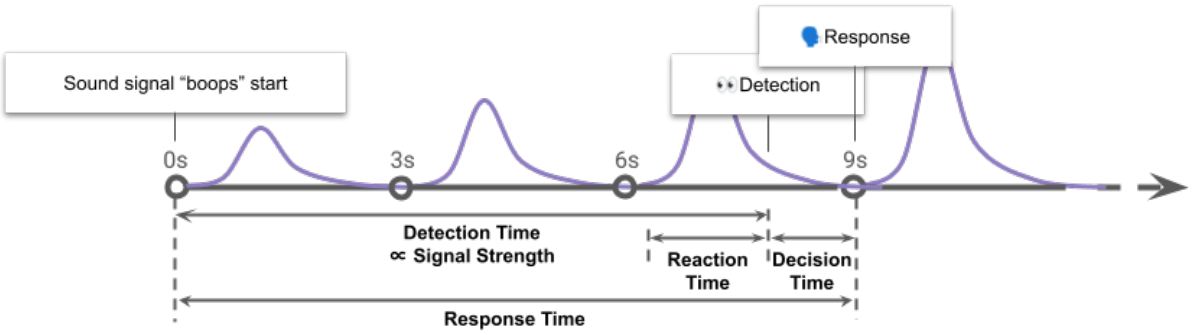
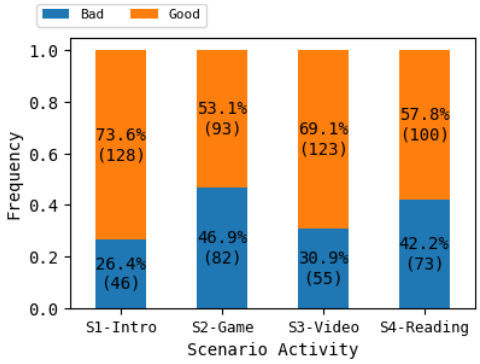


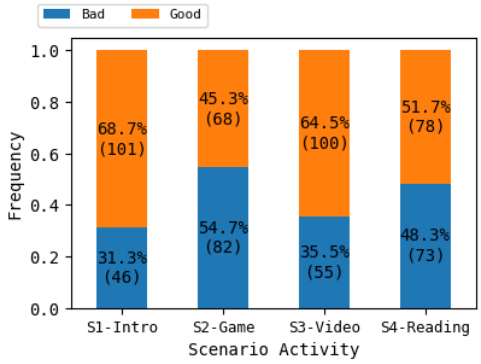
Figure A.4: Timing measurements for each instance of signal detection and response.

B Supplemental Results

B.1 Distribution of Labels among Participants



(a) Aggregated from the labels provided by the participants. 444 (63.4%) out of 700 were "good" labels.



(b) Aggregated from the labels provided by the participants, excluding those who always answered "good". 347 out of 603 (57.5%) "good" labels remained.

Figure B.1: Relative frequency of "good" and "bad" labels in each scenario.

B.2 Simulation and Task Validity

We provide more detailed reports on survey and behavioral measurement data below.

B.2.1 Simulation Immersiveness. The participants' subjective presence ratings on our adapted presence questionnaire averaged 5.64 ± 0.69 on a 1-7 scale. The 11 questions on the questionnaire have a Cronbach's $\alpha \approx 0.74$. The highest ratings are in "How quickly did you adjust to the virtual driving environment experience?" (6.06) and "How well could you identify sounds?" (6.14). Meanwhile, the lowest ratings are for the questions "How much did your experiences in the virtual environment seem consistent with your real-world experiences? " and "How much did the visual aspects of the environment involve you?", both 5.10 on average. Admittedly, we do not have a real-world system to compare to, yet these ratings, at the very least, show that testing UX on our desktop simulator grants good responsiveness and audio authenticity.

B.2.2 Task Load in Different Scenarios. Our quantitative measurements show that the three assigned tasks (Game, Video, and Reading) sufficiently manipulated the participants' level of engagement and cognitive load. On the NASA TLX questionnaire (six questions on a 1-7 scale), both the Game ($M = 28.72$, $SD = 5.22$) and Reading ($M = 24.86$, $SD = 5.46$) induced a higher overall task load on the participant compared to the Video case ($M = 19.91$, $SD = 4.40$) (repeated measures ANOVA: $F(3, 63) = 21.81$, $p < 0.001$; posthoc Tukey's HSD: $p < 0.001$ for Game vs. Video and $p = 0.014$ for Reading vs. Video). Specifically, playing the mobile game and reading short stories created higher mental demand (r.m. ANOVA: $F(3, 63) = 15.32$, $p < 0.001$; posthoc Tukey's HSD: $p < 0.001$ for Game $M = 6.05$, $SD = 1.13$ vs. Video $M = 3.91$, $SD = 1.44$ and $p = 0.008$ for Reading $M = 5.41$, $SD = 1.68$ vs. Video) and time demands (r.m. ANOVA: $F(3, 63) = 25.26$, $p < 0.001$; posthoc Tukey's HSD: $p < 0.001$ for Game $M = 5.81$, $SD = 1.37$ vs. Video $M = 2.32$, $SD = 1.67$ and $p < 0.001$ for Reading $M = 4.41$, $SD = 1.76$ vs. Video). Both Game and Reading produced higher perceived effort as well (r.m. ANOVA: $F(3, 63) = 14.35$, $p < 0.001$; posthoc Tukey's HSD: $p = 0.005$ for Game $M = 5.59$, $SD = 1.33$ vs. Video $M = 4.00$, $SD = 1.45$ and $p = 0.037$ for Reading $M = 5.27$, $SD = 1.64$ vs. Video). The differences between Game and Reading in task load were significant only in terms of the physical (r.m. ANOVA: $F(3, 63) = 12.49$, $p < 0.001$; posthoc Tukey's HSD: $p = 0.01$ for Game $M = 3.36$, $SD = 1.99$ vs. Reading $M = 1.95$, $SD = 1.43$) and time demands (posthoc Tukey's HSD: $p = 0.026$ for Game $M = 5.82$, $SD = 1.37$ vs. Reading $M = 4.41$, $SD = 1.76$), as expected.

B.2.3 Detection and Response Timing. We ran two-way ANOVA on the detection and response timing measurements, assuming that our participants' ability to notice the audio cue was affected by both their familiarity with the signal (training effect) and the tasks we assigned them. Our analyses showed that both the number of days in the simulator and the NDRT scenario had significant effects on detection time (day: $F(3, 691) = 4.02$, $p = 0.008$, scenario: $F(3, 691) = 18.96$, $p < 0.001$), response time (day: $F(3, 691) = 4.39$, $p = 0.005$, scenario: $F(3, 691) = 29.89$, $p < 0.001$), and decision time (day: $F(3, 691) = 6.05$, $p < 0.001$, scenario: $F(3, 691) = 18.96$, $p < 0.001$). There was also an interaction between the day number and the NDRT scenario affecting the three measurements (day-scenario interaction on detection time: $F(9, 691) = 1.86$, $p = 0.054$, on response time: $F(9, 691) = 3.62$, $p < 0.001$, and on decision time: $F(9, 691) = 2.76$, $p = 0.004$). As for the reaction time, we only saw some weak or insignificant influence from the day ($F(9, 691) = 2.76$, $p = 0.075$) and scenario ($F(9, 691) = 2.76$, $p > 0.1$) factors. The participants gave many more immediate responses (0-second decision time) in the Game scenario (56% of the time vs. 29% in Video and 32% in Reading). Meanwhile, the participants took significantly longer time to detect the sound signal when watching the video lectures ($M = 5.97$s, $SD = 3.22$s) (Tukey's HSD $p < 0.001$ compared to both playing mobile game with $M = 4.10$s, $SD = 2.37$s and reading short stories with $M = 3.88$s, $SD = 2.49$s) – a 2-second delay. This is expected as the participants' audio channels were occupied during the video lectures.

B.2.4 Random Object Identification. Participants in the Video scenario successfully identified $M = 1.35$ objects on average ($SD = 1.07$), significantly more than in Game ($M = 0.43$, $SD = 0.79$) (ANOVA: $F(3, 88) = 3.66$, $p = 0.015$; posthoc Tukey's HSD: $p = 0.009$), although not significantly more than Reading ($M = 1.04$, $SD = 0.98$; Tukey's HSD: $p > 0.1$). Only 6 participants missed all objects in the Video scenario, compared to 11, 16, and 8 in Intro, Game, and Reading, respectively. This situational awareness test served as a good indicator of varied visual engagement in different tasks.

B.2.5 Pupil Diameter. Here we would also like to mention the difference in pupil diameters, potentially another indicator of cognitive load levels [11], between the three assigned tasks. Participants playing the mobile game exhibited significantly larger pupil dilation ($M = 4.94$mm, $SD = 0.72$mm) (Tukey HSD $p < 0.001$ compared to both Video with $M = 3.73$mm, $SD = 0.55$mm and Reading with $M = 3.55$, $SD = 0.53$) – the mean pupil diameter is at least 1.1mm greater. However, we could not eliminate the effect of the content brightness on the participants, as the graphics of *Temple Run* were in a slightly darker shade. The pupil dilation in the Video and Reading cases appeared to be similar (less than 0.2mm of difference), despite the reading materials being presented in black text on a white background. Thus, these findings weakly agree with the evidence presented above. In general, our results suggest that the surrogate tasks effectively manipulated the task load and engagement as designed.

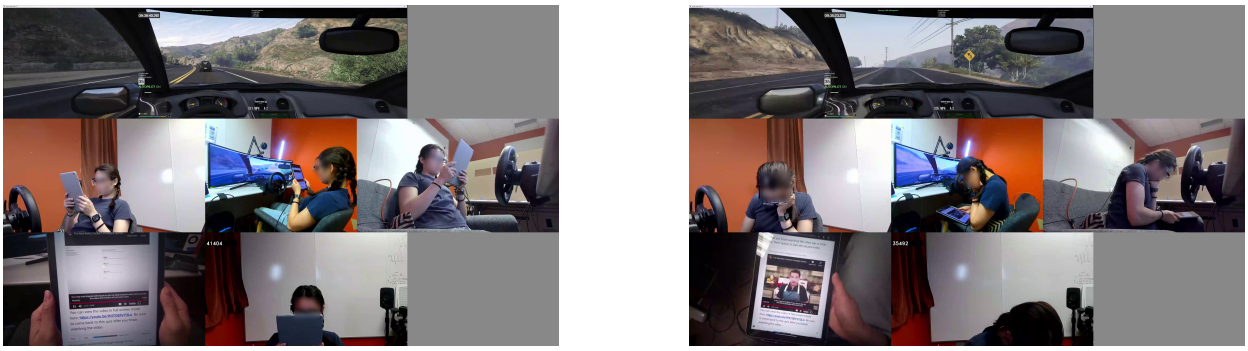
B.3 Interpretations of "Good" Timing

Table B.1 shows examples of how participants interpreted "good" and "bad" moments for each scenario and each category. A few participants (e.g., D113 on the table) did not fall into the same category in different scenarios while most participants remained consistent. Three (3) participants (D101, D109, and D110) always answered "good" throughout the whole study. They explained that they either 1) did not mind the notifications at all thanks to having enough cognitive resources, 2) did not value or entertain the assigned tasks as much, or 3) could pause their tasks and deal with the signals without much annoyance.

Table B.1: Participants' interpretations of what was a "good" or a "bad" moment for them to receive non-emergency notifications during the simulated autonomous car commute. Some participants had different types of interpretations for different NDRTs.

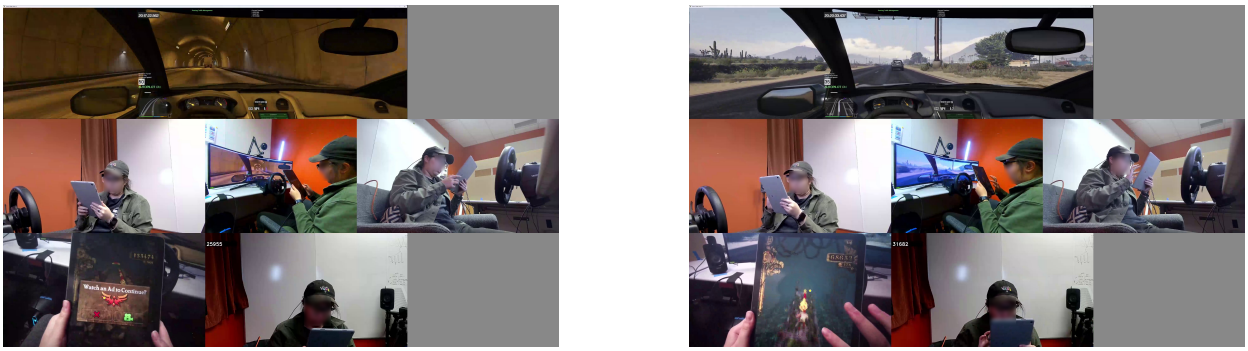
Participant Type	Interpretation Examples			
	1 – Intro	2 – Game	3 – Video	4 – Reading
Always "Good"	"It was all good because I didn't feel there to be a problem if I was interrupted. I reacted a bit slower when I was very involved in my task." —D110	"The game in itself was the most distracting thing so the sound was actually pretty minor really." —D101	"I can watch [the] video and pay attention to [the] notification [at] the same time, so it's all good." —D109	"While reading, I was using more mental energy than while watching a video, but I was free to pause and stop at any time." —D113
Prioritizing NDRTs	"While I worked on a task on my laptop, it was usually a bad moment. While I was taking a break in between tasks, these were good moments." —D116	"If [I was] not actively playing the game, [it] was good." —D114	"Since I was trying to not miss any information in the video, it was always a bad time." —D116	"Mostly bad because I was focused on reading. If I am currently looking at [the] road then good. I mostly only looked up when I heard something else." —D114
Task Content Dependent	"It depends on the task I was doing, I only [called] bad when I was concentrating on reading the text-based article. " —D107	"Good moment means that the sound is not disturbing my game playing, also when the game is not in an intense phase that requires a lot of attention." —D104	"[It was a good moment] when the video was not conveying some sort of important information that I felt was important to the quiz." —D118	"...If there were visuals or much information in the talk, I was focusing on the video and did not like the notifications." —D103
Mental State Dependent	"Just a gut feeling mostly, but usually when I just started to get into the task, or when I wasn't tired of working on mine." —D108	"If I had time to think about it, or if my mind was momentarily drifting anyway." —D122	"I was a bit tired, so it depended on how much energy I had to pay attention to more things... More generally, I answered "good" when I had more mental energy..." —D113	"Whether I was in the flow state. If I was not concentrated and the sound went off, [I'd] say good." —D107

B.4 Additional Figures from Video Data



(a) (b)

Figure B.2: An example of (a) "good" and (b) "bad" moments during a Video scenario. In (a), one of the video lectures was coming to an end. In contrast, in (b) the participant was heavily invested in the informative part of the video.



(a) (b)

Figure B.3: An additional example of (a) "good" and (b) "bad" moments during a Game scenario. In (a), the participant just finished one round of the mobile game. In (b), the participant was in the middle of the game.

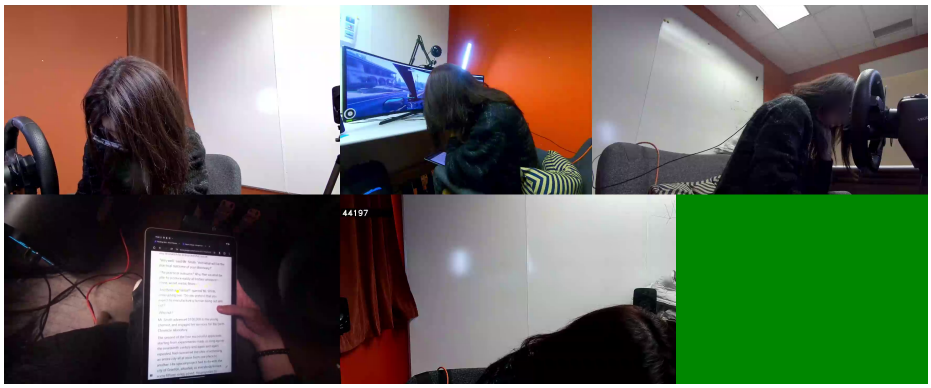


Figure B.4: The participant was deeply focused on the reading material, sitting in a hunched posture with her face occluded by her hand and hair. This would create challenges for face-tracking computer-vision models.

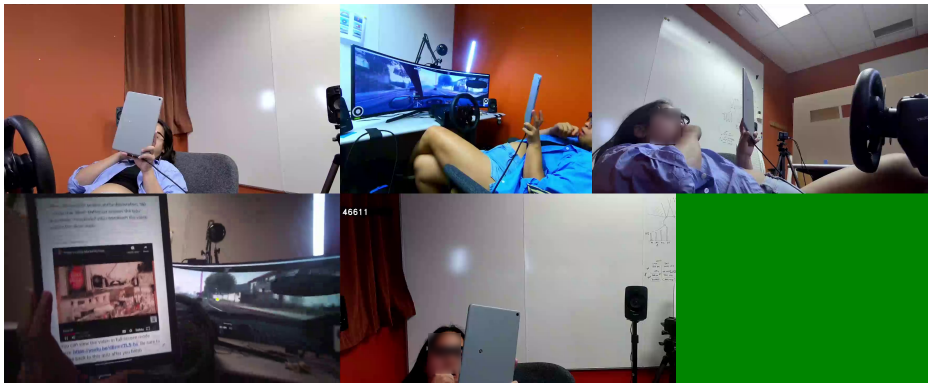


Figure B.5: While watching the video lecture, the participant was reclining. Her face was occluded by the tablet and her right hand. Another challenging situation for face-tracking systems.

C Supplemental Results from Modelling

This section includes figures that provide supplementary information to our report on qualitative observations and modeling.

C.1 Additional Figures for Modeling Reference

The following figures illustrate the x, y, and z components of the eye tracking measurements and vehicle status data.

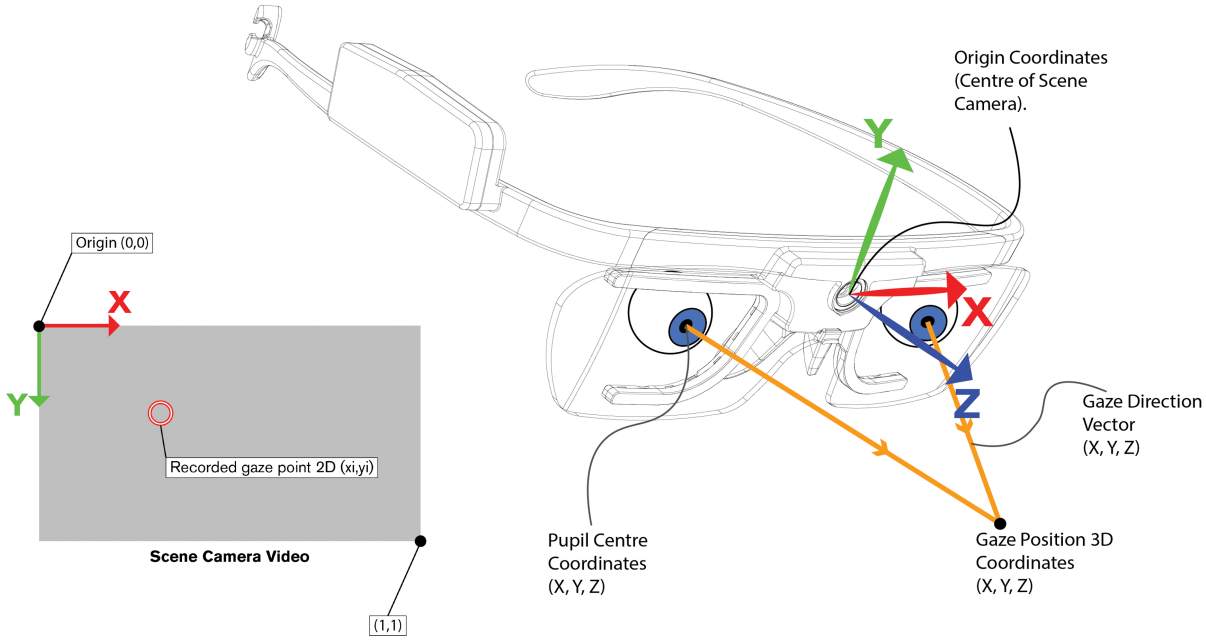


Figure C.1: Gaze measurement coordinates of the Tobii Pro Glasses 2 eye tracker. Retrieved from: https://go.tobii.com/Tobii-Pro-Glasses2-API. All 3-dimensional measurements use the origin at the center of the first-person-view camera. The accelerometer and gyroscope use the same x, y, and z directions. The xy-coordinate for the 2-dimensional gaze position relative to the first-person-view video frame is shown in the bottom left. The origin is positioned at the top left corner of the video frame.

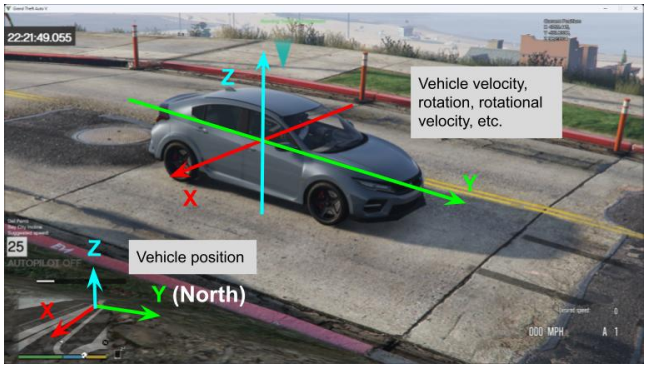


Figure C.2: The coordinate system in the video game *Grand Theft Auto V*. For vehicle motion measurements, the x-axis is in the lateral direction pointing towards the right side of the vehicle; the y-axis is in the longitudinal direction pointing forward; the z-axis points upwards. For the position of the entities on the map, the origin is at the center of the map and the x- and y-axes are oriented eastward and northward, respectively; the z-axis represents the altitude.

C.2 Algorithm Comparison

Table C.1: Comparison between different classifying algorithms in 5-fold cross-validation. All cross-validation runs used the 15s time window, default parameters, and a `random_state` of 1 for all steps that involved random seeds. Some variation could be observed with different random seeds.

Algorithm	Precision	Accuracy	F1 Score	AUC	Note
CatBoost	72.78%	68.83%	0.751578	0.729933	Natively supports categorical features. Symmetrical.
TabPFN	77.39%	69.32%	0.762559	0.712188	The `AutoTabPFNClassifier` was used to automatically search for hyperparameters. Limited tunability.
LightGBM	73.32%	69.81%	0.745956	0.74155	Fastest tree algorithm tested. Asymmetric.
XGBoost	72.90%	68.16%	0.724388	0.720478	Asymmetric.
SVM	58.45%	50.40%	0.505379	0.545891	Does not accept missing values.

C.3 Time Window Sizing

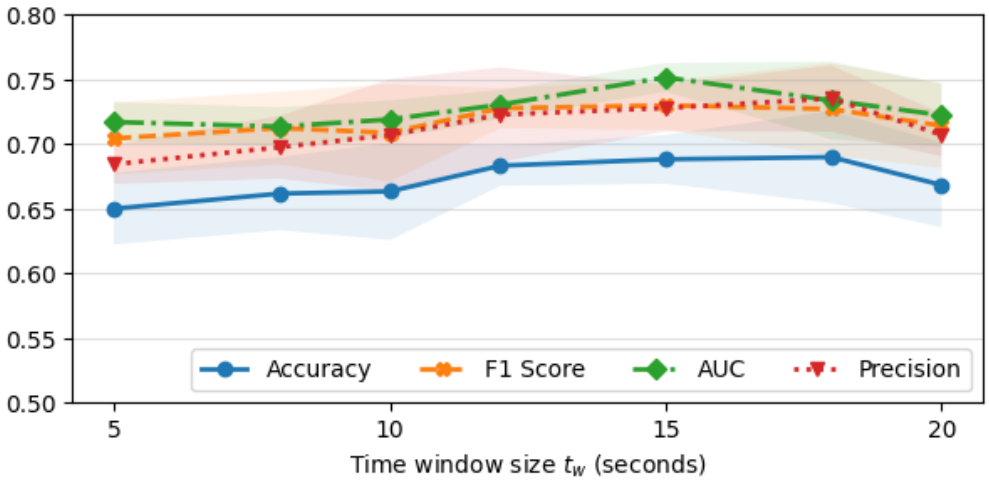


Figure C.3: Comparison between different time window sizes (t_w) in terms of mean accuracy, precision, and AUC. For reference, 57.5% of labels in our full dataset are "good".

C.4 ROC Curves from Tuning and Ablation Studies

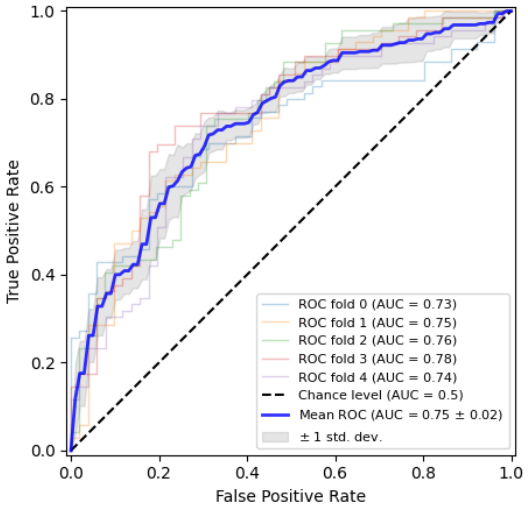
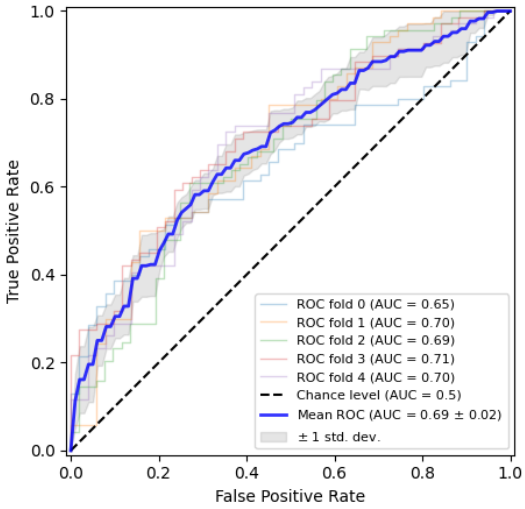
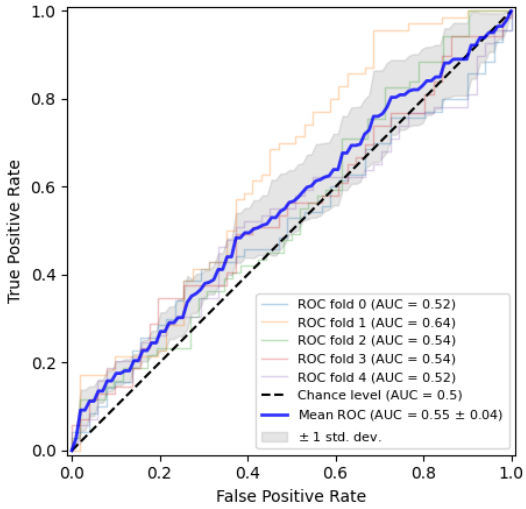


Figure C.4: ROC curves from the stratified 5-fold cross-validations using the default parameters of CatBoost and a `scale_pos_weight` = 0.3.



(a) Without gaze and pupil features from eye tracker data (75.87% precision, 60.86% accuracy, AUC= 0.6913, and F1= 0.5797).



(b) Without any features from eye tracker data, which include eye gaze and head pose information (63.92% precision, 49.75% accuracy, AUC= 0.5527, and F1= 0.3975).

Figure C.5: Comparison of ROC curves from the stratified 5-fold cross-validations after removing sensor data. Both used the default parameters `scale_pos_weight` = 0.3.

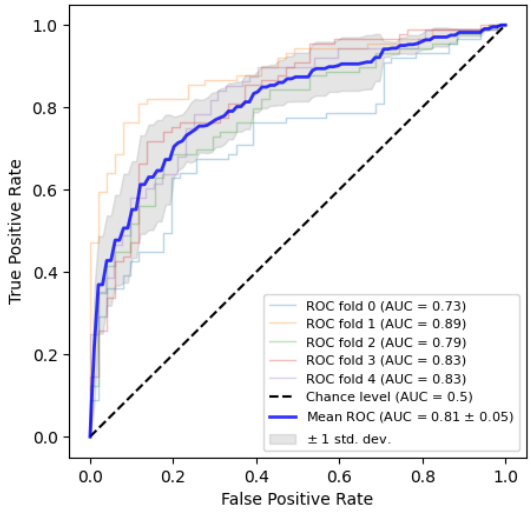


Figure C.6: ROC curves from the stratified 5-fold cross-validations using the default parameters of CatBoost and a `scale_pos_weight` = 0.3. In this case, all available data, including those from participants who always answered "good," were included. The performance may be regarded as optimistic.

Table C.2: Correlations (a) between features and the response time (time needed for the participants to respond to the audio signal by verbally indicating the "good/bad" labels), and (b) between features and the decision time (the time between the detection of the audio signal and the actual "good/bad" response). Only the top 20 features are shown. We also omitted the correlations between features and the reaction time (the time between the most recent "boop" and detection) because the correlations were rather weak and had low statistical significance. The strongest correlation in this case was -0.096* from `tab_acc_y_mean_change`.

(a) Pearson correlations between features and audio signal response time measurements.
*: $p < 0.05$. **: $p < 0.01$.

Feature Name	Corr. w/ t_{response}
tab_touch_count	-0.346**
disp_audio_rms_mean	0.301**
gaze_audio_rms_mean	0.297**
veh_vel_x_mean	0.183**
tab_acc_z_std	-0.168**
veh_rot_y_mean	-0.167**
tab_acc_x_std	-0.167**
gaze_audio_rms_std	0.164**
3d_gaze_pos_x_mean	0.162**
r_pup_cent_y_std	0.154**
r_gaze_dir_z_mean_change	0.154**
r_pup_cent_y_mean	0.154**
r_gaze_dir_z_std	0.148**
veh_nearby_vehicles_count_mean	0.144**
tab_acc_y_mean	-0.143**
r_gaze_dir_y_std	0.141**
disp_audio_rms_std	0.139**
veh_vel_y_mean	0.139**
gaze_pos_y_std	0.136**
l_pup_cent_z_mean	0.127**

(b) Pearson correlations between features and audio signal decision time measurements.
*: $p < 0.05$. **: $p < 0.01$.

Feature Name	Corr. w/ t_{decision}
tab_touch_count	-0.248**
head_gyro_x_mean	0.184**
head_gyro_y_std	0.130**
l_pup_diam_mean	-0.127**
l_gaze_dir_x_std	0.119**
gaze_pos_x_std	0.117**
l_pup_cent_z_mean_change	-0.114**
r_pup_diam_mean	-0.111**
head_accel_z_std	0.103*
head_gyro_x_std	0.103*
tab_acc_x_std	-0.101*
l_pup_cent_x_std	0.101*
r_gaze_dir_x_std	0.098*
tab_acc_z_mean	-0.097*
r_pup_cent_x_mean	-0.096*
time_since_previous	-0.081*
head_accel_y_std	0.080*
head_gyro_y_mean	0.079
3d_gaze_pos_x_mean	0.078
gaze_audio_rms_mean	0.076